

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION  
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : <b>C12Q 1/68</b>		A1	(11) International Publication Number: <b>WO 00/03037</b> (43) International Publication Date: 20 January 2000 (20.01.00)
(21) International Application Number: <b>PCT/US99/15355</b> (22) International Filing Date: 7 July 1999 (07.07.99)		(81) Designated States: CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  <b>Published</b> <i>With international search report.</i>	
(30) Priority Data: 60/092,424 10 July 1998 (10.07.98) US 09/262,127 3 March 1999 (03.03.99) US			
(71) Applicants: THE HOWARD HUGHES MEDICAL INSTITUTE [US/US]; 4000 Jones Bridge Road, Chevy Chase, MD 20815-6789 (US). THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Suite 350, 900 Welch Road, Palo Alto, CA 94304 (US).			
(72) Inventors: BROWN, Patrick, O.; 76 Peter Coutts Circle, Stanford, CA 94305 (US). DIEHN, Maximillian; Apt. 2E, Hulme, Escondido Village, Stanford, CA 94305 (US). EISEN, Michael; 2211 Spaulding Avenue, Berkeley, CA 94703 (US).			
(74) Agent: SHERWOOD, Pamela, J.; Bozicevic, Field & Francis LLP, Suite 200, 285 Hamilton Avenue, Palo Alto, CA 94301 (US).			
(54) Title: SCREENING ARRAYS OF NUCLEOTIDES TO DETERMINE CORRESPONDENCE WITH BOTH SEQUENCE AND PHYSICAL PROPERTIES OF A PROBE			
(57) Abstract <p>Methods are provided for characterizing target polynucleotides on an array. The array is hybridized to probes, where the probes are fractionated and labeled with spectrally distinguishable labels in order to provide information about a physical attribute of the nucleic acid, e.g. length of mRNA, association with ribosomes, sub-cellular localization, etc. Nucleic acids present on the array are scored for hybridization with a probe having particular label characteristics. Based on this information, target nucleic acids are identified that correspond to the desired physical attribute.</p> <p style="text-align: center;"><i>TDS</i></p> <p style="text-align: center;"><i>NOD</i></p>			

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		

**SCREENING ARRAYS OF NUCLEOTIDES TO DETERMINE  
CORRESPONDENCE WITH BOTH SEQUENCE AND PHYSICAL PROPERTIES  
OF A PROBE**

5

**STATEMENT AS TO FEDERALLY SPONSORED RESEARCH**

This invention was made with Government support awarded by National Cancer Institute, grant number HHZ419 and the Howard Hughes Medical Institute. The Government may have certain rights in this invention.

10

**INTRODUCTION**

*Background*

A large number of partial sequences of protein-encoding human, mouse and other animal; plant; and microbial messenger RNAs (mRNAs) have been, and are continuing to be, deposited in databases in the form of Expressed Sequence Tags (ESTs). Thus, an increasingly large number of partial and complete genetic sequences of commercially, scientifically and medically important genes are expected to be known in the near future. It is useful to be able to determine which of these sequences encode polypeptides having a particular functional property, e.g. being secreted, membrane-bound, highly expressed, etc. In addition to being important in the understanding of genetic function, this information is of practical value. For example, secreted and membrane-associated proteins are of particular interest in the development of novel therapeutics and diagnostics.

A tool showing considerable promise for parallel analysis of multiple sequences is the nucleic acid array, reviewed by Ramsay (1998) Nat. Biotech. **16**:40-44, and in a collection of reviews in Nature Genetics (1999) **21**(1 supplement):1-60. These arrays contain dense collections of nucleic acids, either PCR products or polynucleotides, usually of known sequence, that have been either synthesized or printed at fixed spatial locations on suitable substrates, such as nylon filters or glass slides. When labeled DNA or RNA samples are hybridized to the arrays, the abundance of specific sequences in solution can be quantitated based on the fluorescent or radioactive signal intensity at the position of the complementary probe. While recent interest has been

directed toward the use of arrays for monitoring global gene expression, arrays can also be used for rapid detection of sequence variation.

Most of the therapeutics that have been successfully developed by biotechnology companies have been naturally secreted proteins, or modified versions 5 thereof, e.g. insulin, human growth factor, erythropoietin, tissue plasminogen activator, DNase, etc., or agents that bind to cell surface molecules, e.g. herceptin. A large fraction of conventional therapeutics are also directed at cell-surface molecules such as receptors, channels and transporters.

Secreted proteins are also of particular utility in the development of clinical 10 diagnostic tests. Identification of secreted proteins specific to tumors would allow the development of non-invasive blood based assays along the lines of the PSA screen for prostate cancer, that could be used to screen high risk populations. Early detection of the disease should allow early treatment leading to better survival statistics. Furthermore, recent advances in monoclonal antibody therapy call for the identification 15 of new tumor-specific membrane associated markers that could serve as targets for monoclonal antibody therapies. Also, antibodies against secreted mediators of inflammation or other pathological processes, e.g. cytokines, chemokines, etc., could neutralize these offending agents.

Identification of proteins having these functional characteristics will be useful in 20 other fields of biochemical study, including the possible identification of novel molecules involved in inflammation or isolation of new receptors and signaling molecules important in cellular interactions during developmental processes. The partial sequences of the tens of thousands of different cDNAs that are currently available in public databases are certain to include many gene products with various 25 physical attributes. However, these cannot always be identified by their sequences alone.

For example, the sequence of transmembrane regions and signal sequences are helpful, but not decisive in identifying proteins that are membrane bound or secreted. However, the information in EST databases is typically partial sequences 30 corresponding to the 3' region of an mRNA, which may not include the sequences of transmembrane regions or signal peptides that could be recognized using sequence

based rules. Additionally, even when the full-length primary sequence is known, the sequence-prediction methods are very unreliable in predicting whether a given protein is secreted or membrane bound.

5 A high through-put method for determining which gene products fall into defined functional classes based on properties of the protein important class would therefore be of great value.

#### *Relevant literature*

The complete genome sequence of a number of organisms may be found at the  
10 National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>. The availability of sequences of genes of the human genome is discussed in Schuler (1996) *Science* 274:540. The complete sequence of the genome of *S. cerevisiae* is available at several Internet web sites, and is discussed in Goffeau *et al.* (1996) *Science* 274:546.

15 A number of methods are available for creating microarrays of biological samples, such as arrays of DNA samples to be used in DNA hybridization assays. Exemplary are PCT Application Serial No. WO95/35505, published December 28, 1995; U.S. patent no. 5,445,934, issued August 29, 1995; and Drmanac *et al.*, *Science* 260:1649-1652. Yershov *et al.* (1996) *Genetics* 93:4913-4918 describe an alternative  
20 construction of an polynucleotide array. The construction and use of polynucleotide arrays is reviewed by Ramsay (1998) *supra*.

Methods of using high density polynucleotide arrays are known in the art. For example, Milosavljevic *et al.* (1996) *Genomics* 37:77-86 describe DNA sequence recognition by hybridization to short oligomers. The use of arrays for identification of  
25 unknown mutations is proposed by Ginot (1997) *Human Mutation* 10:1-10.

Quantitative monitoring of gene expression patterns with a complementary DNA microarray is described in Schena *et al.* (1995) *Science* 270:467. DeRisi *et al.* (1997) *Science* 270:680-686; DeRisi *et al.* (1999) *Curr Opin Oncol.* 11(1):76-9; and Iyer *et al.* (1999) *Science* 283:83-7, for example, explore gene expression on a genomic scale.  
30 Wodicka *et al.* (1997) *Nat. Biotech.* 15:1-15 perform genome wide expression monitoring in *S. cerevisiae*.

## SUMMARY OF THE INVENTION

Methods are provided for screening nucleic acid arrays, which allow classification of target sequences: by their hybridization to a probe solution of suitable labeled nucleic acids, and by physical attributes associated with that probe. The 5 probes are labeled nucleic acids from a source sample comprising a complex mixture of different nucleic acids. The source sample is fractionated prior to labeling based on a physical attribute, e.g. association with a membrane bound ribosome, association with multiple polysomes, association with specific molecules, e.g. proteins; subcellular localization, etc. Each fraction of interest is labeled. The probe or probes is hybridized 10 to an array, and the labels present on the probe or probes is detected. Nucleic acids present on the array are scored for the presence of the labels. Based on this information, target nucleic acids present on the array are characterized as to their correspondence with a desired physical attribute.

In one embodiment of the invention, the methods provide rapid differentiation 15 between sequences that code for soluble intracellular proteins, and proteins that are either secreted or associated with cellular membrane structures such as the plasma membrane. Probe fractionation for these sequences relies the physical association of membrane bound polysomes with mRNA encoding secreted or membrane bound proteins.

20

## BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a schematic of polysome isolation and DNA microarray hybridization.

25

## DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Methods are provided for characterizing target nucleic acids that are present on an array. Target sequences are identified based on their hybridization to a probe. The probes are labeled in order to provide information about a physical attribute, e.g. length of mRNA, association with ribosomes, sub-cellular localization, etc. The probes are 30 obtained from a source sample that is fractionated based on such a physical attribute. The fractions are then differentially labeled. The probes hybridize to an array, and the

label present on the probes is detected. Nucleic acids present on the array are scored for hybridization with a probe having particular label characteristics. Based on this information, target nucleic acids are identified that correspond to the desired physical attribute.

5       The source nucleic acid may be fractionated to give a single fraction of interest, which is then labeled for use as a probe. The hybridization pattern of the probe is usually compared to the hybridization pattern of a comparison probe labeled with a spectrally resolvable label. The comparison probe may be another fraction from the separation, the unfractionated material, or any other suitable reference sequence. It  
10 will be understood by one of skill in the art that many detectable labels useful in conjugating nucleic acids are available, and can be used to provide combinatorial diversity in the subject methods.

The source nucleic acid may be separated into two or more fractions, where the label is used to distinguish between fractions. For example, a fractionation that yields  
15 two fractions of material may be labeled as follows: fraction (1) with fluorochrome X; and fraction (2) with fluorochrome Y. Analysis of the hybridized array for relative intensities of the X and Y fluorochrome allows identification of a target nucleic acid as corresponding to fraction (1) or (2). Alternatively, a selected fraction may be labeled with fluorochrome X; and a sample of the unfractionated source nucleic acid labeled  
20 with fluorochrome Y.

#### *Source Sample*

The source sample provides the basis for the labeled probe. In a preferred embodiment of the invention, the source sample is mRNA from a cell population, which  
25 is fractionated based on physical properties, e.g. association of the mRNA with sub-cellular organelles or proteins; sub-cellular localization of the mRNA; nuclear transport; size of the mRNA species, etc. The fractionated mRNA may be directly labeled, indirectly labeled, or amplified prior to labeling.

In many instances, the isolation procedure will be performed in conjunction with  
30 the fractionation for a desired physical property, as the association of mRNA with proteins and organelles in the intact cell provides useful information about the encoded

protein. Where the basis for fractionation is length of RNA species, then conventional methods of isolation may be used, where protein associations are not maintained.

Desirably, the source sample is derived from a species where a significant amount of genomic or cDNA sequence information is available. A number of organisms have sufficient sequence information to meet these requirements, including organisms with complete known genome sequences, e.g. *Aquifex aeolicus*; *Archaeoglobus fulgidus*; *Bacillus subtilis*; *Borrelia burgdorferi*; *Escherichia coli*; *Haemophilus influenzae*; *Helicobacter pylori*; *Methanobacterium thermoautotrophicum*; *Methanococcus jannaschii*; *Mycoplasma genitalium*; *Mycoplasma pneumoniae*; *Saccharomyces cerevisiae*; *Synechocystis PCC6803*; and organisms with substantial sequence and mapping information known, e.g. *Arabidopsis thaliana*; *Caenorhabditis elegans*; *Drosophila melanogaster*; *Homo sapiens*; *Leishmania major*; *Mus musculus*; *Oryza sativa*; *Zea mays*, etc.

15

#### *Fractionation*

The source sample is fractionated according to a physical property. Where the source sample is mRNA, the physical property may provide information about the encoded polypeptide. For example, secreted and membrane bound mRNA species are typically associated with membrane bound polysomes. Alternatively, the polysomes may be fractionated by precipitation with antibodies, e.g. pan-specific antibodies that recognize a peptide motif or domain of interest. Subcellular localization, association with specific proteins, nuclear transport, and translational regulation are all properties that can be measured by physical fractionation of mRNAs. Furthermore, by fractionating total mRNA populations on the basis of size using denaturing gel-electrophoresis, followed by labeling of fractions representing different size classes and hybridization to microarrays, the size of the mRNA represented by each unknown gene represented in a microarray can be determined.

In one embodiment of the invention, cytoplasmic and membrane bound polysomes are separated. This can be accomplished via various subcellular fractionation protocols, including ones that are detergent or density gradient centrifugation based. A description of the density gradient based isolation processes

can be found in Mechler (1987) *Meth. Enzymol.* 152:241-247. The mRNA molecules of secreted or membrane-bound proteins are translated by ribosomes that are found in the membrane-bound polysomal fraction.

- Polysome precipitation with antibody species uses a two step fraction, and may
- 5 be performed in combination with the subcellular fractionation. The polysome complexes that are isolated by the above methods may further comprise nascent polypeptide chains encoded by the mRNA. An immunoprecipitation is performed to further fractionate the polysomes into those that bind to an antibody, and those that do not. For example, see Netuschil and Kurth (1980) *J Virol Methods* 1(2):99-112;
- 10 Palmiter *et al.* (1972) *J Biol Chem* 247(10):3296-304. Of particular interest for polysome immunoprecipitation are antibodies that are pan-specific for a particular motif; class of proteins; domain; amino acid residue, e.g. phosphotyrosine, phosphoserine; and the like. Examples of such antibodies may be found in U.S. Patent no. 5,763,198; Choi *et al.* (1997) *Arch Biochem Biophys* 344(1):165-75; Keel *et al.*
- 15 (1995) *Endocrinology* 136(3):1197-204; Cheng *et al.* (1991) *Mol Biochem Parasitol* 49(1):73-82; and others. The use of this specific technique permits identification of classes of polypeptides encoded by genes represented in an array.

- In another fractionation protocol of interest, polysomes are separated by density according to the size of the polysomal complex. The size is correlated with the number
- 20 of ribosomes present on each mRNA molecule, and is indicative of the level of translation of that mRNA, where a greater number of ribosomes are present when translation levels are high. Probes derived from these fractions will therefore identify sequences whose corresponding mRNAs have different levels of translation in the cell type, or in the physiologic condition, of interest. Any suitable method may be used to
- 25 fractionate the polysomes, such as sucrose gradients and the like.

mRNAs may be fractionated according to length on gels, gradients, etc. as known in the art. These fractions then identify sequences in an array based on the length of the mRNAs that they represent.

*Amplification*

The fractionated nucleic acid sample may be labeled directly, or it may first be amplified in order to provide larger amounts, increased signal, etc. mRNA may be amplified by RT-PCR, using reverse transcriptase to form a complementary DNA strand, followed by polymerase chain reaction amplification using primers specific for the subject DNA sequences, or by joining oligonucleotides to the ends of the cDNAs, and amplifying the entire cDNA sample by using primers specific for these added oligonucleotides. Alternatively, mRNA may be amplified by the methods described in U.S. Patent no. 5,545,522, Van Gelder *et al.* Briefly, cDNA is synthesized from a ribonucleic acid (RNA) sequence using a complementary primer linked to an RNA polymerase promoter region. Anti-sense RNA is transcribed from the cDNA by an RNA polymerase capable of binding to the promoter region.

*Labeling Fractions*

A detectable label may be included in such amplification reactions, or may be conjugated to the fractionated nucleic acid after, or in the absence of, amplification reactions. Standard labeling protocols for nucleic acids are described in Sambrook *et al.*; Kambara *et al.* (1988) BioTechnology 6:816-821; Smith *et al.* (1985) Nuc. Acids Res. 13:2399-2412.

Suitable labels include fluorochromes, e.g. fluorescein isothiocyanate (FITC), rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'- dichloro-6-carboxyfluorescein (JOE), 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7- hexachlorofluorescein (HEX), 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6- carboxyrhodamine (TAMRA); etc.

Sets of fluorochromes suitable for multicolor analysis are known in the art. For example, Haddad *et al.* (1998) Hum Genet 103(5):619-25 discuss the number of spectrally distinguishable fluorochromes or fluorochrome combinations, and the simultaneous visualization of probes in 24 different colors. Other fluorochromes are known in the art, for examples see Wiegant *et al.* (1996) J Histochem Cytochem 44(5):525-9; Vaandrager *et al.* (1996) Blood 88(4):1177-82; van Gijlswijk *et al.* (1997) Histochem Cytochem 45(3):375-82; and others.

The label may also be an indirect system, where the amplified DNA is conjugated to biotin, haptens, etc. having a high affinity binding partner, e.g. avidin, specific antibodies, etc., where the binding partner is conjugated to a detectable label.

The label may be conjugated to one or both of the primers in an amplification.

- 5 Alternatively, the pool of nucleotides used in the amplification is labeled, so as to incorporate the label into the amplification product.

A quickly and easily detectable signal is preferred, and fluorescent tagging of the probe is preferred. Other suitable labels include heavy metal labels, magnetic probes, chromogenic labels, e.g. phosphorescent labels, dyes, and fluorophores, 10 spectroscopic labels, enzyme linked labels, radioactive labels, and labeled binding proteins. The label will be selected such that at least two different labels can be simultaneously measured in one hybridization.

- Where multiple fractions are labeled, each fraction will be labeled in such a way that it is distinguished from the other fractions or from a reference nucleic acid sample.
- 15 this can be as simple as labeling two fractions, each with a different fluorochrome. Alternatively, each fraction can be compared to the same common reference by labeling each individual fraction, in turn, with fluor 1, and labeling a reference mixture, which might be a selected fraction, or composed of samples of each fraction, with fluor 2, and then for each fraction hybridizing the mixture of fluor 1-labelled fraction with fluor 20 2-labeled reference mixture to an array, and then measuring the fluorescence ratio of the two fluors at each array element. (see DeRisi *et al.* (1997), *supra*. for examples of the use of a common reference probe ) In this way, quantitative properties, such as mRNA size or translational complex size, can be related to a series of fractions, and identified in an array.
  - 25 In some cases, it may be advantageous for several fractions to be analyzed simultaneously using a more complex labeling scheme, where each fraction will have a distinguishable fluorochrome or combination of fluorochromes. This can be as simple as labeling two fractions, each with a different fluorochrome. Where more fractions are to be used as probes, the labeling scheme may be more complex, where each fraction 30 will have a distinguishable fluorochrome or combination of fluorochromes. In this way,

quantitative properties, such as mRNA size or translational complex size, can be related to a series of fractions, and identified in an array.

### Array

5        The term array as used herein is intended to refer to high density collections of nucleic acid sequences. Arrays may be gridded on a planar surface, or may be attached to particles. High density microarrays of polynucleotides are known in the art and are commercially available.

10      The sequence of polynucleotides on the array will correspond to cDNA or genomic sequences, usually obtained from the species that is the source of the probe. Arrays of interest for the subject methods will generally comprise at least about  $10^2$  different sequences, usually at least about  $10^3$  different sequences, and may comprise  $10^4$ ,  $10^5$ , or more different sequences. This number of sequences may be deposited on a small planar surface of about 1-10 cm<sup>2</sup>. The length of polynucleotide present on  
15      the array is an important factor in how sensitive hybridization will be to the presence of a mismatch. Usually polynucleotides will be at least about 12 nt in length, more usually at least about 15 nt in length, preferably at least about 20 nt in length or longer.

Methods of producing large arrays of polynucleotides are described in U.S. Patent no. 5,134,854 (Pirrung *et al.*), and U.S. Patent no. 5,445,934 (Fodor *et al.*) using  
20      light-directed synthesis techniques. The polynucleotides are synthesized stepwise on a substrate at positionally separate and defined positions. Use of photosensitive blocking reagents allows for defined sequences of synthetic steps over the surface of a matrix pattern. By use of the binary masking strategy, the surface of the substrate can be positioned to generate a desired pattern of regions, each having a defined  
25      sequence polynucleotide synthesized and immobilized thereto.

Alternatively, microarrays are generated by deposition of pre-synthesized polynucleotides onto a solid substrate, for example as described in U.S. Patent no. 5,807,522, issued September 15, 1998. The production of the collection of specific polynucleotides may be produced in at least two different ways. Present technology  
30      certainly allows production of ten nucleotide oligomers on a solid phase or other synthesizing system. See e.g., instrumentation provided by Applied Biosystems, Foster

City, Calif. Alternatively, amplification reactions are used to generate specific fragments of DNA, or clones contained inserts of the desired sequences may be grown by conventional techniques. Once the desired repertoire of possible oligomer sequences of a given length have been synthesized, this collection of reagents may  
5 be individually positionally attached to a substrate. This attachment could be automated in any of a number of ways.

Multiple substrates may be simultaneously exposed to the probe. In this case, each polynucleotide may be attached to a single bead or substrate. The beads may be encoded to indicate the specificity of attached polynucleotide. The probe is then  
10 bound to the whole collection of beads and those beads that have appropriate complementary sequence will hybridize to the probe. A sorting system may be utilized to sort those beads that actually bind the target from those that do not. This may be accomplished by presently available cell sorting devices or a similar apparatus. After the relatively small number of beads which have bound the probe are collected, the  
15 encoding scheme may be read off to determine the specificity. An encoding system may include a magnetic system, a shape encoding system, a color encoding system, or a combination of any of these, or any other encoding system.

#### *Hybridization*

20 The hybridization conditions between probe and target is selected such that the hybridization of the two molecules is both sufficiently specific and sufficiently stable. See Hames and Higgins (1985) Nucleic Acid Hybridisation: A Practical Approach, IRL Press, Oxford. These conditions will be dependent both on the specific sequence and often on the guanine and cytosine (GC) content of the complementary hybrid strands.  
25 The conditions may often be selected to be universally equally stable independent of the specific sequences involved. This typically will make use of a reagent such as an alkylammonium buffer (see Wood et al. (1985) P.N.A.S. 82:1585-1588; and Krupov et al. (1989) FEBS Letters 256:118-122). An alkylammonium buffer tends to-minimize differences in hybridization rate and stability due to GC content. By virtue of the fact  
30 that sequences then hybridize with approximately equal affinity and stability, there is relatively little bias in strength or kinetics of binding for particular sequences.

The different probes may be combined in a single hybridization reaction, or each probe may be applied to a separate array substrate. In a preferred embodiment, two or more differently labeled probes are combined in the hybridization reaction.

5

### *Detection of Signal*

Methods for analyzing the data collected by fluorescence detection are known in the art. Data analysis includes the steps of determining fluorescent intensity as a function of substrate position from the data collected, correcting for background, removing technically deficient data, *i.e.* data deviating from a predetermined statistical distribution, and calculating the relative binding affinity of the targets from the remaining data. Although not necessary, the resulting data may be displayed as an image with the intensity or color in each region varying according to the binding affinity between targets and probes.

Arrays can be scanned to detect hybridization of the labeled probes. Methods and devices for detecting fluorescently marked targets on devices are known in the art. Such detection devices often include a microscope and light source for directing light at a substrate. A photon counter detects fluorescence from the substrate, while an x-y translation stage varies the location of the substrate. A confocal detection device that may be used in the subject methods is described in U.S. Patent no. 5,631,734. A scanning laser microscope is described in Shalon *et al.* (1996) Genome Res. 6:639. A scan, using the appropriate excitation line, is performed for each fluorophore used. The digital images generated from the scan are then combined for subsequent analysis. For any particular array element, the ratio of the fluorescent signal from one probe is compared to the fluorescent signal from the other probe(s), and the relative signal intensity determined. Other devices may use a CCD camera to image the entire array or segments of the array.

The initial data resulting from the detection system is an array of data indicative of fluorescent intensity versus location on the substrate. The data are typically taken over regions substantially smaller than the area in which the polynucleotide was attached. For example, with a "spot" of 500 microns by 500 microns, the data may be taken over regions having dimensions of 5 microns by 5 microns. Within any "spot",

a large number of fluorescence data points may be collected. The detection method provides a positional localization of the region where hybridization has taken place, and this position is then correlated with the corresponding nucleic acid attached to that position.

5

*Corresponding full-length Sequence*

Where the arrayed nucleic acid sequences are obtained from partial cDNA clones (or ESTs), it is desirable to obtain the corresponding full-length sequence. A nucleic acid having a sequence of such an EST, or a portion thereof comprising at least 15 - 20 nucleotides, is used as a hybridization probe to detect the complementary sequence in a cDNA library using probe design methods, cloning methods, and clone selection techniques as known in the art. Libraries of cDNA are made from selected tissues. Preferably, the tissue is the same as the tissue from which the EST or the probe was obtained. The choice of cell type for library construction can be made after the identity of the protein encoded by the gene corresponding to the nucleic acid of the invention is known. This will indicate which tissue and cell types are likely to express the related gene, and thus represent a suitable source for the mRNA for generating the cDNA.

Techniques for producing and probing nucleic acid sequence libraries are described, for example, in Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. The cDNA can be prepared by using primers based on sequence from the partial cDNA, or using poly-T primers.

Members of the library that encompass the complete coding sequence of the native message are obtained. In order to confirm that the entire cDNA has been obtained, RNA protection experiments are performed as follows. Hybridization of a full-length cDNA to an mRNA will protect the RNA from RNase degradation. If the cDNA is not full length, then the portions of the mRNA that are not hybridized will be subject to RNase degradation. This is assayed, as is known in the art, by changes in electrophoretic mobility on polyacrylamide gels, or by detection of released monoribonucleotides. Sambrook et al., Molecular Cloning: A Laboratory Manual, 2nd

Ed., (1989) Cold Spring Harbor Press, Cold Spring Harbor, NY. In order to obtain additional sequences 5' to the end of a partial cDNA, 5' RACE (PCR Protocols: A Guide to Methods and Applications, (1990) Academic Press, Inc.) may be performed.

PCR methods may be used to amplify the members of a cDNA library that comprise the desired insert. The desired insert will contain sequence from the full length cDNA that corresponds to the EST sequence. Such PCR methods include gene trapping and RACE methods. "Rapid amplification of cDNA ends," or RACE, is a PCR method of amplifying cDNAs from a number of different RNAs. The cDNAs are ligated to an oligonucleotide linker, and amplified by PCR using two primers. One primer is based on sequence from the instant nucleic acids, for which full length sequence is desired, and a second primer comprises sequence that hybridizes to the oligonucleotide linker to amplify the cDNA. Another PCR-based method generates full-length cDNA library with anchored ends without needing specific knowledge of the cDNA sequence. The method uses lock-docking primers (I-VI), where one primer, poly TV (I-III) locks over the polyA tail of eukaryotic mRNA producing first strand synthesis and a second primer, polyGH (IV-VI) locks onto the polyC tail added by terminal deoxynucleotidyl transferase (TdT). This method is described in WO 96/40998.

The subject nucleic acid compositions can be used to, for example, produce polypeptides, as probes for the detection of mRNA in biological samples, e.g., extracts of cells, to generate additional copies of the nucleic acids, to generate ribozymes or antisense oligonucleotides, and as single stranded DNA probes or as triple-strand forming oligonucleotides.

For convenience, kits may be supplied which provide the necessary reagents in a convenient form and together. For example kits could be provided that include chips containing an appropriate microarray for the subject to be analyzed, labels and fractionation reagents. Other components such as automated systems for determining and interpreting the hybridization results, software for analyzing the data, or other aids may also be included depending upon the particular protocol which is to be employed.

It is to be understood that this invention is not limited to the particular methodology, protocols, cell lines, animal species or genera, and reagents described,

as such may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

As used herein the singular forms "a", "and", and "the" include plural referents 5 unless the context clearly dictates otherwise. Thus, for example, reference to "a cell" includes a plurality of such cells and reference to "the array" includes reference to one or more arrays and equivalents thereof known to those skilled in the art, and so forth. All technical and scientific terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which this invention belongs unless 10 clearly indicated otherwise.

All publications mentioned herein are incorporated herein by reference for the purpose of describing and disclosing, for example, the cell lines, constructs, and methodologies that are described in the publications which might be used in connection with the presently described invention. The publications discussed above 15 and throughout the text are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the inventors are not entitled to antedate such disclosure by virtue of prior invention.

The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the subject 20 invention, and are not intended to limit the scope of what is regarded as the invention.

Efforts have been made to ensure accuracy with respect to the numbers used (e.g. amounts, temperature, concentrations, etc.) but some experimental errors and deviations should be allowed for. Unless otherwise indicated, parts are parts by weight, molecular weight is average molecular weight, temperature is in degrees 25 centigrade; and pressure is at or near atmospheric.

## EXPERIMENTAL

### Large-Scale Identification of Secreted and Membrane-associated Gene Products using DNA Microarrays

30 We used a modification of the protocol described by Mechler (1987) Methods in Enzymology 152:241-8 to isolate rER-bound polysomes from Jurkat cells, a human

leukemic T cell line, and the protocol described by Stoltenburg *et al.* (1995) *BioTechniques* 18:564-8, to isolate rER-bound polysomes from an exponentially growing culture of *S. cerevisiae* (Figure 1). In each case, we first separated membrane-associated polysomes from those free in the cytoplasm. The Eberwine  
5 RNA amplification protocol (Eberwine, *et al.* 1992. PNAS 89:3010) was used to amplify mRNA sequences present in fractionated polysome samples from Jurkat cells while mRNA samples from yeast cells were analyzed without amplification. We prepared labeled cDNA copies of mRNA from the two fractions using different fluorescent dyes (red for the membrane-bound and green for the free fraction), combined the two  
10 labeled cDNAs, and hybridized the mixture to a DNA microarray. The array used for the Jurkat experiment contained approximately 5500 human cDNAs while the array used in the yeast experiment contained almost every known putative protein-encoding gene in the yeast genome (approximately 6500 open reading frames). Using a modified confocal microscope, we separately scanned at the wavelength appropriate  
15 for each dye and overlaid the two images to create the final composite image. The fluorescent hybridization signals from the two labeled samples at each array element were quantitated using the ScanAlyze program.

The majority of proteins that are translated on the rER should fall into one of three classes: membrane-associated, secreted, or ER/Golgi resident. Of the 200  
20 cDNAs whose transcripts were most enriched in the membrane-bound polysomal fraction from Jurkat cells, 137 represented named genes and 63 corresponded to uncharacterized genes (i.e. ESTs). For 118 of the 137 named genes, published reports that described the subcellular localization of their products could be found (Table 1). 92% of these gene products were reported to be either membrane-  
25 associated, secreted, or ER/Golgi-resident.

Table 1

	Free	Membrane associated	% Membrane-associated	Uncertain	EST
<u>top:</u> 200 genes	10	108	92%	19	63
300 genes	20	154	89%	27	99

- Table 1. Break-down of mRNAs that were most enriched in the membrane-bound fraction of Jurkat cells. Known gene products were categorized as coding for "Free" (cytosolic, cytoskeletal, or nuclear) or "Membrane-associated" (secreted, transmembrane, extracellular, or ER/Golgi-resident). A number of named gene products were not easy to classify based on literature searches and were listed as "Uncertain." ESTs are novel cDNA sequences which do not code for named genes. "% Membrane-associated" refers to the percentage of the sum of the "Free" and "Membrane-associated" categories.
- 5                  The statistical results for the named cDNAs suggest that approximately 90% of the 99 ESTs most enriched in the membrane-bound fraction encode membrane-associated, secreted, or ER/Golgi-resident proteins. These results are from the analysis of a single cell line under a single condition – many genes present on the array could not be assessed because they were not detectably expressed in this cell population. By using an analogous procedure to analyze other tissues and cell lines we believe that it should be possible to classify thousands of unknown genes in a similar way.
- 10                Membrane-bound polysomes were purified from yeast cells by a different subcellular fractionation method. Nevertheless, the mRNAs most enriched in the 20 membrane fraction appeared to encode secreted or membrane-associated proteins. The Yeast Protein Database (<http://quest7.proteome.com/YPDhome.html>) was used as the source of subcellular localization information of the products of defined yeast genes. Of the 300 mRNA species most enriched in the membrane fraction, 9 (7% of the named genes in this set) were reported to encode cytoplasmic or nuclear proteins 25 (YPD categories: cytoplasmic, cytoskeletal or nuclear), while 129 (93% of the named genes in this set) coded for membrane-associated proteins (YPD categories: mitochondrial, vesicles of secretory system, endoplasmic reticulum, Golgi, vacuolar, lipid particles, plasma membrane, peroxisomal, extracellular, or unspecified membrane). The remaining 162 cDNAs encoded known or putative proteins for which 30 YPD lacked subcellular localization information (Table 2).

Table 2

	Free	Membrane-associated	% Membrane-associated	Undefined
<u>top: 200 genes</u>	4	100	96%	96
300 genes	9	129	93%	162
400 genes	20	157	89%	223
500 genes	33	180	85%	287

5 Table 2. Break-down of mRNAs that were most enriched in the membrane-bound fraction of exponentially growing yeast cells. Subcellular localization was obtained from the Yeast Protein Database (YPD) and gene-products were grouped into "Free" (YPD categories: cytosolic, cytoskeletal, or nuclear) and "Membrane-associated" (YPD categories: mitochondrial, vesicles of secretory system, endoplasmic reticulum, Golgi, vacuolar, lipid particles, plasma membrane, peroxisomal, extracellular, or unspecified membrane) classes. "% Membrane-associated" refers to the percentage of the sum of the "Free" and "Membrane-associated" categories.

10

In conclusion, we have found that isolation of mRNA molecules based on their association with membranous structures and analysis of these mRNA molecules using DNA microarrays provides a highly parallel means for identification of known and novel genes that code for membrane-associated and secreted proteins. By combining 15 information on subcellular localization with detailed information on expression patterns (obtained using the same DNA microarrays), we believe that it will be possible to rapidly identify secreted and surface proteins that are associated with processes of biological and medical interest. Some of the proteins identified in this way are likely to provide useful markers for non-invasive medical diagnostic tests, or new therapeutic 20 agents and targets.

#### Materials and Methods:

##### Jurkat mRNA Fractionation:

25 Polysome isolation: Grow up tissue culture cells in roller bottles ( $5 \times 10^8$  cells/gradient). Treat with 50  $\mu$ M cycloheximide for 10' at 37°. Pellet cells in 250 ml centrifuge tubes for 10 min at 2800 RPM (4°). After first round of centrifugation, aspirate supernatant and replace with remaining media from roller bottles. Centrifuge as before. Repeat as necessary. Wash cells 2x with PBS supplemented with 50  $\mu$ M cycloheximide. Count cells and resuspend to  $2.5 \times 10^8$  cells/ml in ice cold hypotonic- 30 lysis buffer medium RBS (50  $\mu$ M cycloheximide, 10 mM KCl, 1.5 mM MgCl<sub>2</sub>, 10 mM

Tris-HCl, pH 7.4). Perform remaining steps in cold room. Swell on ice for 5-10 min. Dounce homogenize cells with 10 strokes of a tight-fitting glass homogenizer or alternatively with 10 passes through a cell cracker. If desired, take a small aliquot for later analysis as total cell lysate. Centrifuge lysate for 2 min at 2000g to pellet nuclei.

- 5 Add 2ml of supernatant to 11ml of 2.5M sucrose TK<sub>150</sub>M (150 mM KCl, 5 mM MgCl<sub>2</sub>, 50 mM Tris-HCl, pH 7.4) and mix well. Construct sucrose step gradient in Ultra-Clear 25x89mm centrifuge tubes (Beckman) for SW-28 ultracentrifuge rotor. First place 4 ml 2.5M sucrose TK<sub>150</sub>M in bottom of tube. Next, carefully layer the 13 ml containing the lysate from onto the 2.5 M sucrose cushion. Finally, successively layer 13 ml 1.98M sucrose TK<sub>150</sub>M and 6 ml 1.3 M sucrose TK<sub>150</sub>M onto the gradient. Balance centrifuge tubes by adding 1.3 M sucrose TK<sub>150</sub>M as needed. Centrifuge for 15 hours at 90,000g. Harvest gradients by puncturing bottom of centrifuge tubes with an 18-gauge needle and collecting 1 ml fractions in 1.5 ml microcentrifuge tubes. Measure absorbance of fractions at 260 nm to determine presence of nucleic acid. Free ribosomes and free mRNA will be present in the load zone while membrane-associated ribosomes and mRNA will be at the interface between the 1.98 M and 1.3 M sucrose steps. Alternatively, the membrane fraction can be isolated with a pipet from the top of the gradient.
- 10
- 15

- 20 **RNA Isolation and Microarray Hybridization:** Separately pool load zone fractions (Free RNA) and 1.98M/1.3M interface fractions (Membrane-associated RNA) and isolate total RNA from each with *TRIZOL* Reagent (Life Technologies, Inc.) At this point can either further isolate polyA<sup>+</sup> RNA or use total RNA as input for labeling reaction. If using total RNA, use 50-100µgRNA in normal labeling reaction. If desired, amplify mRNA using the Eberwine RNA amplification protocol (Eberwine, et al., PNAS 89:3010, 1992). Perform DNA microarray hybridization, labeling free RNA with one dye and membrane-associated RNA with a second dye.
- 25

**S. cerevisiae mRNA Fractionation:**

- 30 **Polysome Isolation:** Grow 2 L culture of yeast cells to log phase. Centrifuge at 6000xg, 10 min, 4°C. Wash 1x in DEPC distilled water. Resuspend pellet in ~2ml

- buffer I (20 mM HEPES [pH 7.4], 100 mM potassium acetate, 2 mM magnesium acetate). Transfer liquid nitrogen to a pre-cooled mortar. Drip cells into the mortar to shock freeze. Immediately crush cells using a pre-cooled pestle and transfer to a 15 ml conical. If desired, add more buffer I after careful thawing of the crushed cells.
- 5    Centrifuge at 1000xg, 5 min to remove cell debris. Set aside some supernatant to use as total RNA. Centrifuge remaining supernatant at 10,000xg for 8 min to pellet the membrane bound polysomes (MBP). Save supernatant since it contains free polysomes (FP). Wash and pellet in buffer I. Resuspend in buffer I and add sodium deoxycholate to a final concentration of 0.2% and incubate on ice for 5 min. Add
- 10   Tween 20 (polyoxyethylenesorbitan monolaurate) to a final concentration of 0.5%. Tween 20 is purchased as a liquid stock solution. Incubate on ice for another 5 min. (The detergents release MBP from the rER.) Centrifuge supernatants from steps 7) and 9) at 27,000xg for 10 min to remove membranes. Supernatants contain FP and MBP respectively.
- 15         RNA Isolation and Microarray Hybridization: Isolate mRNA using **TRIZOL** Reagent (Life Technologies, Inc.). Perform DNA microarray hybridization, labeling FP mRNA with one dye and MBP mRNA with a second dye.

**WHAT IS CLAIMED IS:**

1. A method of screening an polynucleotide array, the method comprising:  
fractionating a source sample comprising a complex mixture of nucleic acids into  
fractions based on a physical attribute;  
5 labeling a fraction from said fractionating to provide probes;  
labeling a comparison probe with a spectrally resolvable label;  
hybridizing a polynucleotide array with said probes;  
detecting the presence of said spectrally distinguishable labels on said array;  
wherein the presence said labels indicates that the hybridizing polynucleotide  
10 corresponds to a source nucleic acid comprising said physical attribute.
  
2. The method of Claim 1, wherein said comparison probe is a second  
fraction from said fractionating.
  
- 15 3. The method of Claim 1, wherein said comparison probe is an  
unfractionated source sample.
  
4. The method according to Claim 1, wherein said source sample comprises  
mRNA species.  
20
5. The method according to Claim 4, wherein said physical property is  
association with a polypeptide or organelle.
  
6. The method according to Claim 5, wherein said polypeptide is a nascent  
25 polypeptide chain encoded by said mRNA.
  
7. The method according to Claim 5, wherein said organelle is a membrane-  
bound ribosome.
  
- 30 8. The method according to Claim 4, wherein said physical property is size.

9. The method according to 8, wherein said physical property is size of the polysome-mRNA complex.

10. The method according to Claim 6, wherein said physical property is  
5 length of the mRNA species.

11. The method according to Claim 4, wherein said labeling comprises  
amplification of said fractionated mRNA.

10 12. The method according to Claim 11, wherein said labeling comprises  
incorporation of label during amplification.

13. The method according to Claim 4, wherein said polynucleotide array  
comprises cDNA sequences.

15 14. The method according to Claim 4, wherein said polynucleotide array  
comprises cloned or amplified segments of genomic DNA.

15. The method according to Claim 1, wherein said polynucleotide array  
20 comprises a planar array of at least about  $10^2$  different sequences/cm<sup>2</sup>.

16. A method of screening an polynucleotide array for the presence of  
polynucleotide sequences corresponding to secreted or membrane-bound  
polypeptides, the method comprising:

25 fractionating a source sample of mRNA into fractions based on the association  
with membrane-bound ribosomes;

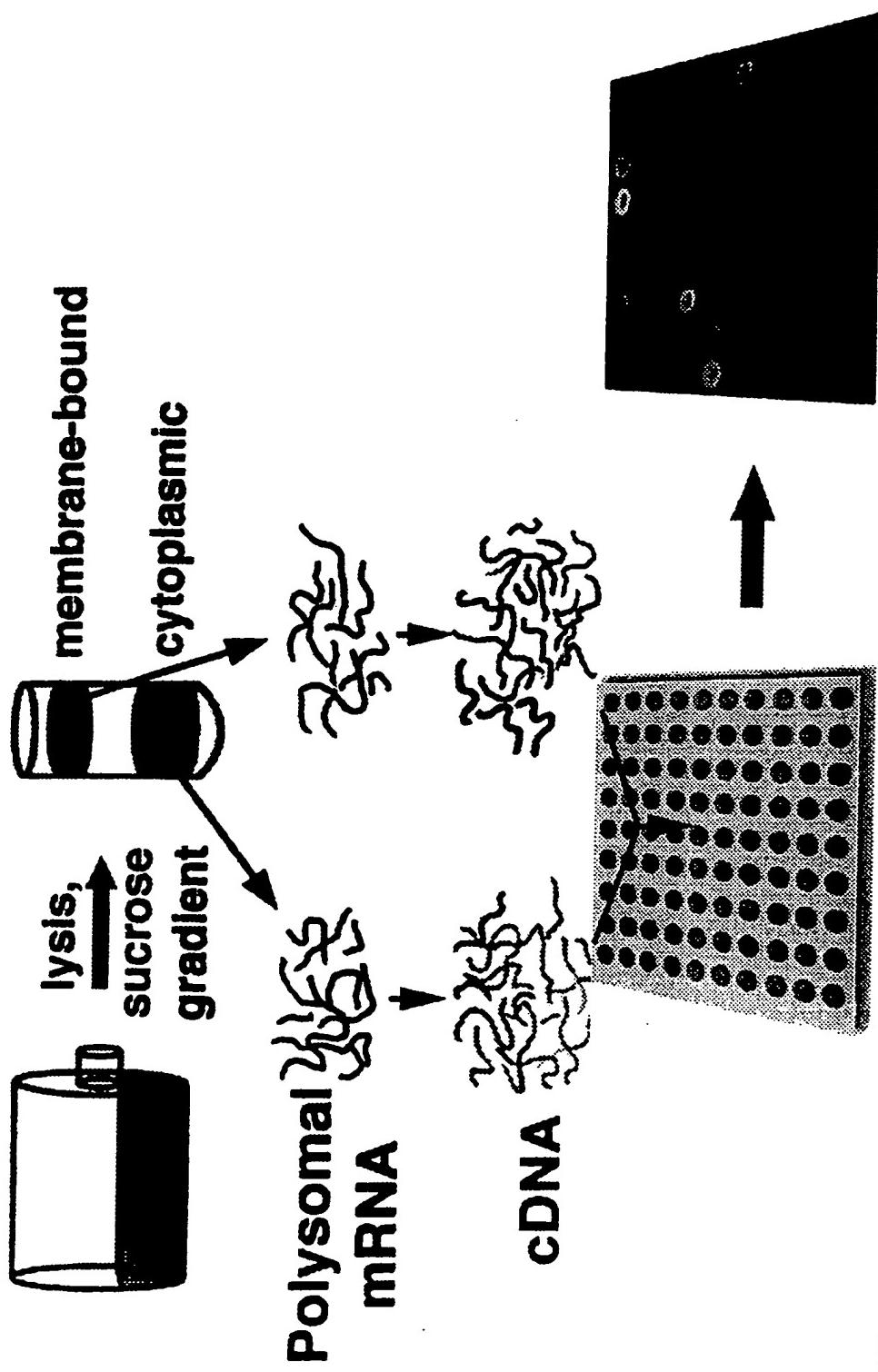
labeling each of said fractions with a different fluorochrome or predefined ratio  
of two fluorochromes, such that each fraction is spectrally distinguishable from every  
other fraction, to provide probes;

30 hybridizing a polynucleotide array comprising polynucleotide sequences with  
said probes;

quantitating said spectrally distinguishable labels at each position corresponding to a discrete polynucleotide sequence on said array;

wherein the presence said labels indicates that the hybridizing polynucleotide corresponds to an mRNA encoding a secreted or membrane associated polypeptide.

1/1



**Figure 1.** Schematic representation of Jurkat polysome isolation and DNA microarray hybridization.

## INTERNATIONAL SEARCH REPORT

International application No. PCT/US99/15355
---

**A. CLASSIFICATION OF SUBJECT MATTER**

IPC(6) :C12Q 1/68

US CL :435/6

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 912, 7.92, 969, 973; 536/24.3, 24.33, 26.6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, GENBANK, EMBASE

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X, P	US 5,800,992 A (FODOR et al) 01 September 1998, See entire document.	1-16
X	US 5,445,934 A (FODOR et al) 29 August 1995, col.30, lines 40-45.	15
X	SCHENA. M. et al. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray Science. 20 October 1995. Vol. 270. pages 467-470, see entire document.	1-16

Further documents are listed in the continuation of Box C.

See patent family annex.

* Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* document defining the general state of the art which is not considered to be of particular relevance		
*B* earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel if it cannot be considered to involve an inventive step when the document is taken alone
*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*O* document referring to an oral disclosure, use, exhibition or other means	"&"	document member of the same patent family
*P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

25 AUGUST 1999

Date of mailing of the international search report

27 OCT 1999

Name and mailing address of the ISA/US  
 Commissioner of Patents and Trademarks  
 Box PCT  
 Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

ARUN CHARRABARTI

Telephone No. (703) 306-5818